Creole languages and genes: the case of São Tomé and Príncipe[1]

Tjerk Hagemeijer[2], Jorge Rocha[3]

1. INTRODUCTION

This article focuses on the gene-language connection between the Portuguese-related Gulf of Guinea creole-speaking populations in São Tomé and Príncipe. The Gulf of Guinea creoles constitute a young language family of four languages spoken on three islands: Santome (ST) and Angolar (AN) on the island of São Tomé; Principense (PR) on Príncipe; and Fa d'Ambo (FA) on Annobón. The latter island, which integrates Equatorial Guinea, is not included in our genetic case-study because its population has not yet been sampled. Figure 1 shows the islands in the Gulf of Guinea.

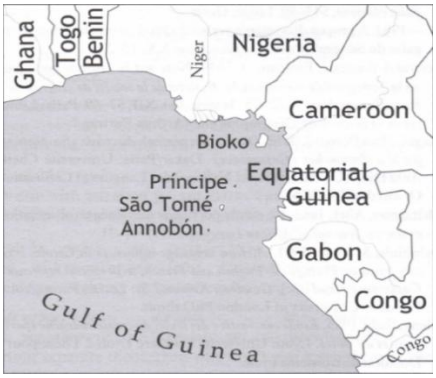

Figure 1. Map of the Gulf of Guinea.[4]

[1] This work was supported by the Portuguese Foundation of Science and Technology (FCT) as part of the project *The origins and development of creole societies in the Gulf of Guinea: An interdisciplinary study* (PTDC/CLE-LIN/111494/2009).
[2] Centro de Linguística da Universidade de Lisboa and Faculdade de Letras, Universidade de Lisboa
[3] CIBIO/InBIO: Research Centre in Biodiversity and Genetic Resources and Departamento de Biologia, Faculdade de Ciências, Universidade do Porto
[4] Map drawn from Michaelis *et al.* (2013: 50)

The peopling of the São Tomé and Príncipe archipelago started in the late 15th century when the island of São Tomé was settled by the Portuguese and slaves from the African continent. This led to the formation of a creole language which branched into four different languages in the 16th century. We argue that these languages show a linguistic founder effect that can be traced back to Nigeria[5], the oldest slave trade area, and that linguistic differentiation corresponds to genetic differentiation in the case of the AN speaking people but not in the case of PR speakers.

The paper is structured as follows. Section 2 provides a short history of the Gulf of Guinea islands focusing on the early settlement; section 3 discusses the lexical and grammatical contribution of African languages to the Gulf of Guinea creoles; section 4 analyzes the genetic structuring of three of the four creole-speaking societies of São Tomé and Príncipe; section 5 concludes.

2. HISTORY OF THE GULF OF GUINEA ISLANDS

The Gulf of Guinea islands were discovered in 1470/1 and permanent, large-scale settlement of São Tomé, the main island, occurred in 1493, following a failed attempt in 1485. In only a few decades, the initial homestead society on the island had already given rise to a sugar-producing plantation society that reached its peak in the second half of the 16th century and then quickly declined in the first half of the 17th century, when the plantation owners left massively for Brazil. The better quality of the sugar in the New World, the threat posed by other European nations who started navigating the Gulf of Guinea, and the attacks on the Santomean plantation system by runaway slaves are among the main factors underlying the exodus from the islands

There is substantial historical evidence that the large majority of the slaves that arrived on São Tomé during the homestead period (approx. 1493-1520) were taken from the Niger delta area (e.g. Caldeira 2008; Ryder 1969; Teixeira da Mota 1976; Thornton 1992; Vogt 1973), where the Portuguese had established diplomatic relations with the ancient kingdom of Benin. With the rise of the plantation society in the 1510s and the labor intensive sugar production process, the number of Bantu slaves began to increase significantly and rapidly became dominant, whereas the Benin trade quickly started fading and was fully cut off in the mid-16th century. Ryder (1969: 65), for instance, concludes that in the period 1525–27 the number of Benin slaves imported to São Tomé was probably at most one-sixth of the total number of arrivals.

Although there is substantial evidence that the African population quickly outnumbered the Portuguese in the new settlement, many question marks still remain regarding its precise demographics. The manuscript written by Valentim Fernandes (1506–1510) (Seibert 2007a), based on second-hand information, mentions that in 1493 the first settlers took 2000 Jewish orphans from Seville

[5] We use the broad label Nigeria to refer to the Niger delta and, in particular, to the area to its west whose center was the old kingdom of Benin.

with them to São Tome in order to settle the island. Only 600 of these children would still be alive in 1506, arguably due to the harsh conditions in the tropics. He further states that in 1506, a decade before the labor intensive sugar cycle started, the island is inhabited by 1.000 *moradores* 'residents', 2.000 slaves and that 5.000 to 6.000 slaves were temporarily on the island. The group of *moradores*, most of which of Portuguese origin, consisted mainly of convicts, but also included a few clergy members, soldiers, other men, free women, and presumably the 600 orphans.[6]

Miscegenation is often referred in the early documents and represents an official and deliberate policy to settle the islands issued, for the first time, by King John II, who established that every Portuguese man in the new colony is entitled to an African female slave. Two decades later, in 1515, King Manuel I, on request of the *moradores*, granted manumission to the children of mixed offspring and two year later, in 1517, the male slaves who had arrived with the first settlers also became free (Caldeira 1999). This led to the rise of a new social class, the *filhos da terra* 'children of the land' or Forros, from the Portuguese designation *forro* 'free slave'[7], a name which is still used for their language. By the end of the 18th century, the Forros represented 95% of the free population, which made up approximately 30% of the total population (Neves & Ceita, 2004).

After the decline of the sugar cycle in the first half of the 17th century, São Tomé maintained a small role as an *entrepôt* in the Atlantic slave trade until the 19th century, but the Santomean social network underwent another major upheaval when the coffee and cacao production on São Tomé and Príncipe started blossoming in the 19th century. Since this period coincided with the abolishment of slavery, provoking a labor crisis in the 1870s, the colonial power came up with a strategy that relied on the recruitment of indentured laborers (the so-called *contratados* or *serviçais*) from other Portuguese colonies, in particular Angola, Cape Verde, and Mozambique in order to replace the former slaves.[8] The demographic effect of the recruitment of the indentured labor force was such that the large number of arrivals quickly doubled the original population, as shown in Table 1. The descendants of the indentured laborers are known as the Tonga's (e.g. Rougé 1992).

---

[6] Seibert (2007a) critically discusses the information provided by Valentim Fernandes with respect to the Jewish children.

[7] The word *forro* is related to *carta de alforria* 'letter of manumission'.

[8] Between 1875 and 1900, the contract laborers were predominantly recruited in Angola, but in 1879 a small labor force came from Gabon and West-African regions. Recruitment in Cape Verde and Mozambique starts mainly in the early 20th century (Seibert, 2006). The incoming contract workers between 1902 and 1958 are as the following: 80.500 Angolans; 68.782 Mozambicans; 34.668 Cape Verdeans (Seibert, 2006:50).

Table 1. Evolution of the population of São Tomé and Príncipe (1807-1950).[9]

Year	Natives	Portuguese[10]	Contract laborers	Total
1807				11.767
1827				12.713
1843				12.753
1860		151		10.433
1870		449		18.017
1875		741		29.441
1900	18.128	1.012	18.033	42.103
1921	19.196	998	38.697	59.055
1940	31.036	995	28.459	60.490
1950	34.947	1.152	24.060	60.159

The history of São Tomé, the central island, is intrinsically linked to that of the much smaller islands of Príncipe and Annobón. The former island, where PR is spoken, was donated by royal decree in 1500, but, according to Alvaro de Caminha's 1499 will (Albuquerque 1989), at that time the island was already inhabited by people from São Tomé who were sent over to escape the famine. The island of Annobón, where FA is spoken, was donated by royal decree in 1503 but the evidence strongly suggests that permanent settlement only took place in the mid-16th century (Hagemeijer & Zamora 2016 and references therein). The subsequent history of these two islands is quite different, however. The tiny island of Annobón (17,5 km^2) quickly became isolated from São Tomé because it lacked economic and strategic interest. In the late 18th century, the island became a Spanish possession. Príncipe, being a much larger island (136 km^2), maintained a strong connection to São Tomé throughout its history and was granted certain privileges in the early period. During the period of Carneiro's contract (1514-1518), for instance, Principe was granted the monopoly of the Portuguese trade with the kingdom of Benin (e.g. Ladhams 2003). Príncipe went through the sugar, coffee and cacao cycles at a much smaller scale than São Tomé, but the impact of this period is currently still visible in the large population of Cape Verdean descendants.

In addition to these two islands which are inhabited by self-identifying creole societies, another group, the Angolares, inhabits the southern and north-western coastal regions of São Tome and are generally assumed to be descendants of maroon slaves that fled from the island's initial settlement and the sugar plantations in the 16th century (Caldeira 2004; Ferraz 1974; Lorenzino

[9] Data adapted from Nascimento (2000).

[10] The number of Portuguese (or 'white') has been consistently low. Lucas (2015), for example, shows that the number of whites in the period 1758-1822 in the archipelago is at most 1,4%, mostly men, whereas the number of blacks in the same period is always over 95% and 2 to 3% are of mixed race.

1998; Seibert 2007b). Runaway slaves are well documented since the very early stages of the settlement and became a serious threat to the settlement and the plantations from the 1530s on. Santos (1996: 78) found that as many as 684 slaves fled into the jungle between 1514 and 1527, which is well before the peak of the plantation system in 1560/70.

3. The origins of the Gulf of Guinea creole languages

The peopling of the Gulf of Guinea islands described above had far-reaching linguistic consequences. Language contact between Portuguese and continental African languages brought to São Tomé by the slave population during the homestead and plantation stages resulted in the birth of a Portuguese-related creole language. This new language, created by the African population as a means for interethnic communication, resulted from naturalistic second language acquisition with limited access to Portuguese, the lexifier language. It seems reasonable to assume that this creole language went through a stage of pidgnization and expansion before nativization by children transformed it into a full-fledged language. This first creole (the proto-creole of the Gulf of Guinea[11]) was taken to the islands of Príncipe and Annobón and also became the language of the Angolares. The continuation in time and space of the proto-language is known as Santome (also Lungwa Santome, Forro or Creole of São Tomé); the creoles of Príncipe and Annobón are known as Principense (or Lung'Ie) and Fa d'Ambô, respectively; Angolar (also Lunga Ngola or Ngola) is the language of the Angolares. These four languages are currently still spoken but PR in particular is severely endangered.[12]

Despite the lack of mutual intelligibility, the contemporary creoles exhibit a great amount of shared lexical and linguistic features that spread from the proto-language.[13] Ferraz (1979) shows that many of the features found in ST (and, by analogy, the GGCs in general) reveal a strong impact of African continental languages. In his work, he pinpoints Edo, the language of the old kingdom of Benin (Nigeria), which belongs to the larger Edoid family, and Kikongo (western Bantu) as the main African strata. Ferraz did not, however, establish a chronological difference between these two African contributions. However, in the light of the early history described in the previous section, linguistic features that trace back to Edo(id) (and a more general Nigerian typology) that are attested in the GGCs are expected to be related to the homestead period, when the slave trade almost exclusively targeted this area; Bantu-derived features in the GGCs, on the other hand, must have entered the creole(s) slightly later, during

[11] Schang (2000) uses the label 'pré-creole' in the same sense.

[12] See Michaelis *et al.* (2013) for a sociolinguistic overview and a short linguistic description of the GGCs.

[13] The oldest written records of these creoles date back to the late 19th century and in-depth and well-informed descriptions of these languages only became available from second half of the 20th century, which means that reconstructing features of the proto-language has to be undertaken from the contemporary languages.

the plantation society, when the Bantu-speaking areas played a prominent role in the labor supplies to São Tomé.

3.1. Lexicon

The main portion of the lexicon of the GGCs is of course derived from Portuguese but compared to other Portuguese-related creoles, such as the Upper Guinea creoles, the amount of African lexicon is quite substantial, possibly 5 to 10% of the overall lexicon and even more in the case of Angolar (Lorenzino 1998). It is within the African lexicon that we find the most striking lexical differences between these languages. The African lexicon in PR is almost exclusively Edoid-derived (Maurer 2009); ST and FA show a relatively proportional mix of substrate-derived lexical items from Edo and Kikongo (Ferraz 1979; Granda 1985); AN exhibits an unusual high portion of western Bantu lexicon, in particular from Kimbundu (Lorenzino 1998; Maurer 1992, 1995) but also a small number of Edo-derived items. Table 2 presents a few examples of Edoid-derived words that exhibit cognates in the GGCs.

Table 2. Shared Edoid lexicon in the GGCs.

ST	**PR**	**AN**	**FA**	**Edoid**	**meaning**
idu	*idu*	*iru*	*idu*	Edo *ìrù*	louse
budu	*ubudu*	*buru*	*budu*	Emai[14] *údò*, *úkpúdò*, *òβíúdò*	stone
ôbô	*ôvyô*	*ôbô*	*ôgô*	Edo *égbó* Emai *úgbó*	wilderness, forest, jungle
bô	*ba*	*bô*	*bô*	Edo *βòó*	where is/are
a	*a*	*a*	*a*	Edo *à*	impersonal pronoun
ubwê	*igbê*	*ôngê*	*ôguê*	Edo *ègbé*	body

These shared Edoid items not only often belong to the basic vocabulary (e.g. the word for ‘louse’ and ‘stone’) but also include functional lexicon, such as interrogative *bô/ba*, impersonal pronoun *a*, and the forms for ‘body’, a word that that is also used in body-reflexive constructions in the GGCs (and in Edoid languages). Bantu-derived core vocabulary, on the other hand, occurs almost exclusively in AN, i.e. Bantu-related cognates across the four GGCs are highly uncommon. Table 3 presents a few examples of the lexical specificity of AN.

14 Emai is a north-central Edoid language, closely related to Edo.

Table 3. Core lexicon of Bantu origin in Angolar.

ST	**PR**	**AN**	**FA**	**Etymology**	**meaning**
pixi	*pexi*	*kikiê*	*pixi*	Portuguese *peixe* Kimbundu *kikêle*	fish
nôtxi	*unôtxi*	*n'thuku*	*nôtxi*	Portuguese *noite* Kimbundu *usuku*	night
xinku	*xinku*	*tano*	*xinku*	Portuguese *cinco* Kimbundu *tanu*	five

Another contrast with the Edoid-derived lexicon is that the Bantu lexicon in the GGCs is typically related to open word classes, in particular nouns. These abbreviated lexical facts therefore provide a first indication that the language(s) from the oldest slave trade area, in particular Edo, constitute an older layer that was more intricately interwoven with the lexifier.

3.2. Phonology and syntax

In this section we address a number of grammatical features of the GGCs and their putative relation with the two relevant slave trade areas (Nigeria, Congo/Angola). Similarly to the findings with respect to the lexicon in the previous section, it will be shown that the Nigerian typology left a strong imprint on the GGCs. Although most languages from both areas above belong to the Benue-Congo branch of the Niger-Congo family, they are typologically quite distinct. Despite greater variation in peripheral areas, especially in the Northwest, the Bantu languages form a closely-related family of agglutinating languages which share a great amount of lexical and grammatical features. The most relevant Bantu languages for the GGCs are from Guthrie's zone H, which includes the Kimbundu and Kikongo clusters. Several typological features of Nigerian languages are also found in a large sub-Saharan strip that includes languages from West, Central and, to a lesser extent, East Africa. In early work on African languages, these languages were labeled the Sudan languages and thought of as genetically related (Westermann 1911), but nowadays generally considered a typological area (e.g. Güldemann 2008).
In his work on ST, Ferraz noted that (1979: 54), '(…) an overriding aspect of ST phonology is that it is African-based rather than Portuguese-based'. Some of the segmental features or phonological processes that can be related to the African continent and which occur in one or more GGCs are prenasalization, labial velars, implosives, interdentals, open syllable structure, reduced ATR harmony system, depalatalization and palatalization rules and certain sandhi rules (Ferraz 1974, 1975, 1979; Hagemeijer 2009; 2011). Some of these features, such as CV syllable structure, are attested in the relevant Nigerian and Bantu languages and can therefore not be conclusively used to tell these areas apart. A number of features, such as prenasalization and (de)palatalization rules and possibly also interdentals, appear to be indicative of contact with Bantu languages, being

mostly restricted to ST and/or AN. Had these features spread from the proto-GGC, we would expect them to occur in PR and FA as well. Although more research is required, implosives, labial velars, and a reduced ATR harmony system occur in all or can be reconstructed for the proto-GGC (Hagemeijer 2011). These features are widespread in the Macro-Sudan belt and are absent from the relevant Bantu languages (e.g. Clements & Rialland 2008).

The syntax of the GGCs appears to be more homogenous than the lexicon and phonology, providing additional evidence for the importance of the Niger delta area in the making of the proto-GGC. Although the GGCs display several word order patterns that are not attested in Portuguese, such as N-POSS, N-DEM, ADJ-ADV, (NEG) V NEG, or V NP NP (double object constructions), these patterns, which must have spread from the proto-GGC, could be related either to the Nigerian or to Western Bantu languages. Following the line of reasoning we used for the lexicon and phonology, we need to look for shared syntactic features that are distinct from Portuguese, on the one hand, and exclusive to one of the two African continental areas. Examples of this type are widespread verb serialization, body-reflexives, plural-marking with a lexical element corresponding to 3p/pl, prenominal diminutives derived from 'child' (child house=small house), and a final locative interrogative marker used to question NPs only ('where is X'). These are among the structures that are found in many Nigerian languages but not in the Western Bantu languages (e.g. Dimmendaal 2001; Hagemeijer 2011; Parkvall 2000).

3.3. Summary

It was shown above that different regions of the African continent contributed to the lexical and grammatical features of the GGCs. These features, however, entered the creoles at different stages of their history. We assume that features that occur in all or most of the GGCs or that can be reconstructed spread from the proto-GGC, whereas isolated features are more likely to be cases of innovation. Under this assumption, the shared African lexicon and grammatical features generally reflect the typology of the languages of the Nigeria. Given the available evidence, we can further pinpoint Edo or the Edoid family in general as a main contributor, confirming the importance of the work by Ferraz (1979), who was the first scholar to establish several lexical and grammatical parallels between this language and ST. Hence we conclude that the contact language that arose on the island of São Tomé and gave rise to the proto-GGC resulted primarily from the contact between Portuguese and languages from Nigeria. This linguistic founder effect shows that the populations from Nigeria, who were dominant during the homestead period, are the main creators of the new proto-language that split into four different varieties during the 16th century. In this scenario, the contribution of Western Bantu languages is one of secondary contact that arguably set in at a point of time when the proto-language was already stabilizing. The presence of Bantu lexicon and grammatical features, in particular in the domain of phonology, is particularly visible in ST and AN, indicating a strong presence of Bantu-speaking slaves on the island of São Tomé during the

plantation stage. Príncipe arguably did not undergo the same massive impact of Bantu language speakers and PR was therefore able to retain more Nigerian features than the other GGCs. Finally, the case of the GGCs, with a quite abrupt shift in the slave origins in the transition from the homestead to the plantation stages, also shows that creolization in the Gulf of Guinea must have taken place in a relatively short time span.

4. GENETIC STUDIES ON SÃO TOMÉ AND PRÍNCIPE

4.1 Gene language co-evolution

The evolutionary drive that is at the heart of the conceptual link between genetics and linguistics is rooted in the fact that both genes and languages can be vertically transmitted between generations and are affected by population history (Cavalli-Sforza 2000). As much as distributions of languages provide the raw data for linguistic interpretations, the starting point for any population genetic analysis is a set of allele frequencies in different populations across space. Both types of distributions are endpoints of interactions between multiple evolutionary factors that need to be disentangled in order to recover history. Moreover, when intermarriage tends to occur essentially within linguistic groups, language differences may contribute to genetic isolation and be a causal factor of genetic differentiation, either directly or through correlated cultural traits.

Of course, linguistic replacement and lateral transfer will decouple biological microevolution from linguistic changes. Since these discrepancies may be quite common, the association between genes and languages cannot simply be assumed *a priori* and must be tested in every particular case (Barbujani 1991). However, even discordance between different sets of data will provide insights about the influence of historical factors on the current patterns of genetic and non-genetic aspects of human variation.

In this section, we discuss the relationship between linguistic and genetic variation in São Tomé and Príncipe by focusing on genetic findings that may provide a framework for interpreting linguistic diversity in the insular microcosm of the archipelago.

4.2. External population sources

Given that the islands of São Tomé and Príncipe were uninhabited, it is logical to assume that an important component of the current bio-cultural variation in the archipelago was influenced by the relative contributions of external population sources. In order to evaluate the genetic impact of these exogenous contributions, it is important to analyze the demographic input of different areas of origin of slaves from the African mainland, as well as the levels of admixture with the Portuguese settlers.

4.2.1 Origins of slave settlers

To study the origins of slave settlers, Tomás *et al*. 2002 analyzed specific variants of the hemoglobin gene (called hemoglobin S haplotypes), which are particularly useful to assess the origins of migrants with diverse regional African ancestries due to their high levels of geographic segregation in major areas of slave recruitment. Based on a global sample from São Tomé, without assigning individuals to language groups, these authors found evidence for similar contributions of regions from Central-West Africa (Ghana, Benin, Nigeria) and West-Central Africa (Congo and Angola) to the island gene pool.

This trend has been confirmed by Coelho *et al*. (2008), who additionally found that Central-West African and West-Central African haplotypes were equally represented in AN and non-AN speaking groups from the island of São Tomé (see below). Moreover, the lineage composition of the maternally transmitted mitochondrial DNA (mtDNA) diversity, which has levels of geographic segregation comparable to those of the hemoglobin S haplotypes, also seems to reflect the major contributions of those two broad geographic regions to the settlement of the island (Coelho *et al*. 2008). Interestingly, some mtDNA lineages that are known to have a highly localized distribution in the African mainland, and may have originated from the area of the Congo Basin, were found to be virtually restricted to AN speakers (Coelho *et al*. 2008). Our recent analysis of mtDNA lineages from the island of Príncipe also points to similar contributions of Central-West African and West-Central African genetic among individuals who reported to have a PR speaking maternal grandmother (unpublished results).

The relatively high proportion of the Central-West African component uncovered by genetic studies is somewhat surprising, given the available historical evidence for a more important contribution of the Congo-Angola area as a source of slaves to São Tomé, after a marked decline in the Benin trade around 1520 (cf. section 2). One possible explanation for this discrepancy is that a fraction of the slaves from Congo-Angola were subsequently re-exported without leaving a significant contribution to the present genetic makeup of the archipelago. Alternatively, the survival rates of Congo-Angola slaves may have been lower than those of Nigeria, because it seems that especially the former had to endure the harsh conditions of the plantations while the latter were more frequently domestic slaves. In any case, these results are based on the analysis of a limited number of genetic markers. For a more robust conclusion about the contribution of major African areas to the genetic profile of São Tomé and Príncipe, we are currently characterizing this population with a much bigger number of genetic polymorphisms, covering about 2% of the human genome (in preparation). Our preliminary results confirm that Central-West African and West-Central African are the main contributors for the gene pool of São Tomé and Príncipe. Moreover, we found that Central-West African genetic components are slightly more represented in descendants of PR speakers than among ST speakers (40% vs. 30%).

4.2.2 Admixture between Africans and Europeans

By using a set of informative genetic markers from different regions of the genome, Tomás *et al*. (2002) estimated that Europeans (mainly Portuguese) only contributed to about 10% of the present genetic makeup of São Tomé. Coelho *et al*. (2008) have additionally shown that the impact of European admixture in São Tomé is mostly restricted to non-AN speaking groups. Furthermore, the analysis of the paternally inherited Y chromosome and the maternally inherited mtDNA showed that virtually all mtDNA sequences from São Tomé had an African origin, while 15-25% of Y chromosome lineages could be traced to Europe, implying that that admixture was largely due to interbreeding between European males and African women (Trovoada *et al.* 2007; Coelho *et al.* 2008). Finally, no genetic components attributable to a Jewish ancestry could be found, in spite of early references to a large number of Jewish children that were taken to the island of São Tomé by the first settlers (cf. section 2).
Overall, the Portuguese genetic contribution may be considered disproportionately low compared to the European contribution of 43% calculated in the Cape Verde archipelago, which had an initial settling process similar to São Tomé and Príncipe, but never became a true plantation society (Beleza *et al*. 2012).

4.3. Genetic differentiation within the archipelago

4.3.1 São Tomé

Besides the broad contribution of different external sources, historical processes that occurred *in situ*, after the settlement phase, are also likely to have shaped the current genetic profile of São Tomé and Príncipe.
To evaluate how these factors interplayed to promote genetic differentiation among different population groups within the archipelago, Coelho *et al.* (2008) used a study design that attempted to interpret the genetic structure of the island of São Tomé without relying on predefined linguistic or geographical categories. By coupling a transect sampling strategy with a clustering approach in which each individual could be treated as a basic sampling unit, genetic groups could be delimited without relying on non-genetic criteria.
The major finding of this analysis, based on genetic polymorphisms from different regions of the genome, was that the genetic structure of São Tomé is essentially determined by two clusters: one of the clusters grouped most individuals sampled in villages where AN is the predominant autochthonous language; the other cluster grouped individuals from communities living in former plantations that descended from the 19 and 20th century indentured laborers (the Tongas), as well as from ST speaking communities. This extreme genetic differentiation of the AN speaking community is fully confirmed by our recent results based on the genome-wide characterization of genetic diversity (see Section 4.2.1).

Additional analyses of mtDNA and Y chromosome variation have shown that the AN speaking cluster had a pronounced reduction in genetic diversity, indicating that significantly low population sizes might have been a major determinant of the differentiation of this group. This is especially evident in the Y-chromosome, in which a single modal haplotype represented as much as 60% of sampled Angolar lineages. This observation clearly indicates that Angolares carry the imprint of a remarkable founder effect, i.e. a disproportionate representation in the current population of the genes of a single ancestor, or of a group of closely related ancestors. Angolar mtDNA maternal lineages, although less variable than non-Angolar lineages, were found to be more diverse than Y-chromosome variants, implying that population size reduction was more severe in males than in females.
Trovoada *et al*. (2003; 2004) also observed a genetic peculiarity of the Angolares and a lack of differentiation between Tongas and ST speakers in an earlier study of mtDNA and Y chromosome variation, among individuals that self-reported to belong to these groups. However these studies did not detect the marked reduction in Y chromosome diversity observed by Coelho *et al.* (2008), probably because their sampling design was based on predefined categories, being less effective in excluding recent admixture.
Taken together, the results from Coelho *et al.* (2008) show that the AN speaking group might have originated through the flight of a group of patrilineally related rebel slaves who established secondary contacts with the rest of the island mostly through restricted, female-mediated gene flow. In this setting, it is likely that the unique Bantu features found in AN constitute evidence of the original language spoken by fugitive founders who adopted a version of the autochthonous São Tomé creole at a later stage, retaining features of their original language. This scenario is clearly at odds with the perspective that AN speakers descend from a conglomerate of groups with diverse origins that was formed by the assimilation of successive waves of slave runaways.
The identification of a specific African origin for the founders of the Angolar cluster is still not possible with the available genetic evidence. As discussed above, the mtDNA lineages found in this cluster probably derived from Central-West Africa non-Bantu regions and West-Central African regions. The predominant Y chromosome lineage, on the other hand, is likely to have derived from a broad area of West-Central Africa, but this area cannot be further narrowed down. One can speculate, however, that in case of a Bantu origin, the non-Bantu mtDNA component was obtained through female mediated admixture with other population groups from São Tomé.

4.3.2 Príncipe

The genetic studies on the autochthonous population of Príncipe are much less developed than in São Tomé. This situation reflects the rarefaction of descendants from the original African settlers of the island, most probably due to the mortality associated with a major outbreak of sleeping sickness (human African trypanosomiasis) around 1900, which was followed the massive immigration of indentured laborers from the Cape Verde islands (Nascimento 2003).

In this context, finding genetic lineages that can be associated with PR speakers is more similar to actively tracing the segregation of a rare phenotypic trait (like a genetic disease) than to the usual comparisons of unrelated individuals from populations speaking different languages, as in the case of São Tomé.
Interestingly, we found recently that genetic variants affording resistance to human sleeping sickness are more frequent among the descendants of PR speakers than in other regions of Africa, except Nigeria (Pinto *et al.* 2016). However, a much more thorough characterization of this trait in whole archipelago would be required in order to understand if these findings reflect the historic link between Príncipe's settlers and Nigeria, or a selective event favoring the carriers of the resistance variants during the epidemic.
We were also able to identify and characterize 18 Y-chromosome and 42 mtDNA unrelated lineages descending from presumptive PR speaking paternal grandfathers and maternal grandmothers, respectively (unpublished results). We found that the PR associated Y-chromosome lineages from Príncipe are strongly interconnected with other lineages from the island of São Tomé and do not cluster in a distinct group of related sequences, suggesting that, unlike the Angolares, there were no strong male founder effects among PR speakers. Moreover, due to the lack of geographic segregation across Bantu and non-Bantu speaking areas, no clear-cut conclusions could be drawn about the origins of Y-chromosome lineages on the African mainland.
Likewise, the results from the analysis of mtDNA variation showed that PR speakers have a lineage composition that is very similar to that of ST speakers from Príncipe. Taken together these results reveal the lack of a clear association between linguistic and genetic variation on the island of Príncipe.

5. Conclusion

In this brief case-study on the origins and development of the creole societies in the Gulf of Guinea we have shown that an interdisciplinary approach using history, linguistics and genetics contributes to answer old questions and to raise new ones. Departing away from previous claims, such as Ferraz (1979), we argued that the Nigerian and Bantu layers found in the GGCs correspond to different historical periods, i.e. the homestead vs. plantation society. Genetic studies on São Tomé confirm the importance of the Nigerian genotype, which represents at least half of the African contribution. This is somewhat unexpected considering the dominance of the Bantu slave trade to São Tomé over the centuries and leads to the conclusion that Nigerian (freed) slaves and their descendants are historically a stable and demographically important component of the islands population.
The micro-cosmos of São Tomé and Príncipe also shows that linguistic differentiation may or not correspond to genetic differentiation within a relatively shallow time-depth. The Angolares are linguistically and genetically distinct from the remaining population of São Tomé and therefore constitute an example of gene-language co-evolution; the Principense-speakers, on the other hand, are only distinctive from the main population of São Tomé by means of their language, not

genetically. These findings lead to new interpretations of how these creole societies came about and evolved. The Angolares' strong male founder effect related to Bantu-speaking areas is not compatible with maroonage during the homestead society, when the Bantu impact was extremely low or absent, nor with successive waves of male slaves of diverse provenance at later stages. The more diversified female Angolares' lineages, which suggest they were creole women, may ultimately explain why the Angolares speak a creole language and not a Bantu language. The absence of genetic differentiation between the PR and ST speaking groups shows that the PR speaking society was arguably renewed as the result of migration between islands, without fusion of the two languages. It can be excluded that the PR speaking society resulted from a unique historical founder event.

Of course many questions remain. For instance, is the Angolar strong male founder effect compatible with Angolar's unique lexical layer from Kimbundu or are these traits derived from different sources? And how does the Annobón creole society, which appears to have been far more isolated historically, fit genetically into the broader picture of the Gulf of Guinea creole societies? We are hopeful that time will tell.

REFERENCES

Albuquerque, Luis de. 1989. *A Ilha de São Tomé nos Séculos XV e XVI*, Lisboa: Alfa.

Barbujani, Guido. What do languages tell us about human microevolution? *Trends in Ecology and Evolution* 6: 151-156.

Beleza, Sandra *et al*. 2012. The admixture structure and genetic variation of the archipelago of Cape Verde and its implications for admixture mapping studies. *PLoS One* 7:e51103.

Caldeira, Arlindo. 1999. *Mulheres, sexualidade e casamento em São Tomé e Príncipe (séculos XV-XVIII)*. Lisboa: Edições Cosmos.

Caldeira, Arlindo. 2004. Rebeliao e outras formas de resistencia a escravatura nas ilhas do Golfo da Guine (seculos XVI–XVIII). *Studia Africana* 7: 101–136.

Caldeira, Arlindo. 2008. Tráfico de escravos e conflitualidade: o arquipelágo de São Tome e Príncipe e o reino do Congo durante o século XVI. *Ciências & Letras* 44: 55-76.

Cavalli-Sforza, L. Luca. 2000. *Genes, peoples and languages*. New York: North Point Press.

Clements, George. N. & Rialland, Annie. 2008. Africa as a phonological area. In Bernd Heine & Derek Nurse (eds.), *A linguistic geography of Africa*, 36–85. Cambridge: Cambridge University Press.

Coelho, Margarida *et al*. 2008. Human microevolution and the Atlantic slave trade: a case study from São Tome (Gulf of Guinea). *Current Anthropology* 49: 134–143.

Dimmendaal, Gerrit. 2001. Areal diffusion versus genetic inheritance: An African perspective, 358-392. In Alexandra Aikhenvald & R.M.W. Dixon (eds.), *Areal diffusion and genetic inheritance*. Oxford: Oxford University Press.
Ferraz, Luiz Ivens. 1974. A linguistic appraisal of Angolar. In *Memoriam Antonio Jorge Dias*, vol. 2, 177-186. Lisbon: Instituto de Alta Cultura/Junta de Investigações do Ultramar.
Ferraz, Luiz Ivens 1975. African influences on Principense creole. Offprint of *Miscelânea luso-africana*, 153-64.
Ferraz, Luiz Ivens 1979. *The creole of São Tomé.* Johannesburg: Witwatersrand University Press.
Granda, German de. 1985. Las retenciones lexicas africanas en el criollo portugues de Annobon y sus implicaciones sociohistoricas. In German de Granda, *Estudios de lingüistica afro-romanica*, 195–206. Valladolid: Universidad de Valladodid.
Güldemann, Tom. The Macro-Sudan belt: towards identifying a linguistic area in northern sub-Saharan Africa. In Bernd Heine & Derek Nurse (eds.), *A linguistic geography of Africa*, 151–185. Cambridge: Cambridge University Press.
Hagemeijer, Tjerk. 2009. Initial vowel agglutination in the Gulf of Guinea creoles. In Enoch Aboh & Norval Smith (eds.), *Complex processes in new languages*, 29-50. Amsterdam, Philadelphia: John Benjamins.
Hagemeijer, Tjerk. 2011. The Gulf of Guinea creoles: genetic and typological relations. *Journal of Pidgin and Creole Languages* 26(1): 111-154.
Hagemeijer, Tjerk & Armando Zamora. 2016. Fa d'Ambô: Past and present. *International Journal of the Sociology of Language*, 239, 193-209.
Ladhams, John. 2003 *The formation of the Portuguese plantation creoles*. Ph.D. thesis, University of Westminster.
Lorenzino, Gerardo. 1998. *The Angolar creole Portuguese of São Tomé: its grammar and sociolinguistic history*. Ph.D. thesis, City University of New York.
Lucas, Patrícia. 2015. The demography of São Tomé and Príncipe (1758-1822): Preliminary approaches to an insular slave society. In J. P. Oliveira e Costa (org.), *Anais de História de Além-Mar XVI*, (pp. 51-78). Lisboa: CHAM, Universidade Nova de Lisboa.
Maurer, Philippe. 1992. L'apport lexical bantou en Angolar. *Afrikanische Arbeitspapiere* 29. 163-174.
Maurer, Philippe. 1995. *L'Angolar: un créole afro-portugaise parlé à São Tomé*. Hamburg: Helmut Buske Verlag.
Maurer, Philippe. 2009. *Principense – Grammar, texts, and vocabulary of the Afro-Portuguese creole of the Island of Príncipe*. London: Battlebridge Publications.
Michaelis, Susanne; Maurer, Philippe; Haspelmath, Martin; Huber, Magnus (eds.), *The survey of pidgin and creole languages, Vol. II. Portuguese-based, Spanish-based, and French-based languagues.* Oxford: Oxford University Press.
Nascimento, Augusto. 2000. *Relações de poder e quotidiano nas roças de S. Tomé e Príncipe. De finais de oitocentos a meados do presente século*. Doctoral Dissertation, Universidade Nova de Lisboa.
Nascimento, Augusto. 2003. O sul da diáspora: cabo-verdianos em plantações de S**.** Tomé e Príncipe e Moçambique / Augusto Nascimento. Praia: Presidência da República de Cabo Verde.

Neves, Carlos Agostinho & Ceita, Maria Nazaré. 2004. *História de S. Tomé e Príncipe. Breve Síntese*. São Tomé.
Parkyall, Mikael. 2000. *Out of Africa*. London: Battlebridge Publications.
Pinto, Joana Correia *et al*. 2016. Food and pathogen adaptations in the Angolan Namib desert: tracing the spread of lactase persistence and human African trypanosomiasis resistance into southwestern Africa. *American Journal of Physical Anthropology*. 161: 436-447.
Rougé Rougé, Jean-Louis. 1992. Les langues des Tonga. In Ernesto d'Andrade & Alain Kihm (eds.), *Actas do colóquio sobre crioulos de base lexical portuguesa*, 171-176. Lisboa: Colibri.
Ryder, Alan. 1969. *Benin and the Europeans 1485-1897*. London: Longman.
Schang, Emmanuel. 2000. *L'émergence des créoles portugais du Golfe de Guinée*. Doctoral Dissertation, Université de Nancy 2.
Seibert, Gerhard. 2006. Comrades, clients and cousins. Colonialism, socialism and democratization in São Tomé and Príncipe. Leiden: Brill.
Seibert, Gerhard. 2007a. 500 years of the manuscript of Valentim Fernandes, a Moravian book printer in Lisbon. In Beata Elzbieta Cieszynska (ed.), *Iberian and Slavonic cultures: contact and comparison*, 79–88. Lisbon: CompaRes.
Seibert, Gerhard. 2007b. Angolares of Sao Tome island. In Philip Havik & Malyn Newitt (eds.), *Creole societies in the Portuguese colonial empire*, 105–126. Bristol: Bristol University Press.
Teixeira da Mota, Avelino. 1976. Alguns aspectos da colonização e do comércio marítimo dos Portugueses na África Ocidental nos séculos XV e XVI. Lisboa: Junta de Investigações Científicas do Ultramar.
Thornton, John. 1992. *Africa and Africans in the making of the Atlantic world, 1400-1680*. Cambridge: Cambridge University Press.
Tomás, Gil *et al*. 2002. The peopling of Sao Tome (Gulf of Guinea): Origins of slave settlers and admixture with the Portuguese. *Human Biology* 74: 397–411.
Trovoada, Maria de Jesus *et al*. 2003. Evidence for population sub-structuring in Sao Tome and Principe as inferred from Y-chromosome STR analysis. *Annals of Human Genetics* 65: 271–283.
Trovoada, Maria de Jesus *et al*. 2004.Pattern of mtDNA variation in three populations from São Tomé e Príncipe. *Annals of Human Genetics* 68: 40–54.
Trovoada, Maria de Jesus *et al*. 2007. Dissecting the genetic history of São Tomé e Príncipe: a new window from Y-chromosome biallelic markers. *Annals of Human Genetics* 71: 77–85.
Vogt, John. 1973. The Early São Tomé-Príncipe Slave Trade with Mina: 1500-1540. *International Journal of African Studies*, VI, 3: 453-467.
Westermann, Diedrich. 1911. *Die Sudansprachen: Eine sprachvergleichende Studie*. Abhandlungen des Hamburgischen Kolonialinstituts, 3. Hamburg: L. Friederichsen.